THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech

Citation for published version:
Barra-Chicote, R, Yamagishi, J, King, S, Montero, JM & Macias-Guarasa, J 2010, 'Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech', *Speech Communication*, vol. 52, no. 5, pp. 394-404. https://doi.org/10.1016/j.specom.2009.12.007

Digital Object Identifier (DOI):
10.1016/j.specom.2009.12.007

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Speech Communication

Publisher Rights Statement:
© Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. / Barra-Chicote, Roberto; Yamagishi, Junichi; King, Simon; Montero, Juan Manuel; Macias-Guarasa, Javier.
In: Speech Communication, Vol. 52, No. 5, 01.05.2010, p. 394-404.
Research output: Contribution to journal › Article

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 04. May. 2023

Analysis of Statistical Parametric and Unit Selection Speech Synthesis Systems Applied to Emotional Speech

Roberto Barra-Chicote[b,1], Junichi Yamagishi[a], Simon King[a], Juan Manuel Montero[b], Javier Macias-Guarasa[c]

[a] *The Centre for Speech Technology Research, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB United Kingdom*
[b] *Grupo de Tecnologia del Habla, Universidad Politecnica de Madrid, ETSI Telecomunicación, Ciudad Universitaria s/n, 28040 Madrid, Spain*
[c] *Department of Electronics, University of Alcala, Ctra. de Madrid-Barcelona, Km. 33,600, 28805-Alcalá de Henares (Madrid), Spain*

Abstract

We have applied two state-of-the-art speech synthesis techniques (unit selection and HMM-based synthesis) to the synthesis of emotional speech. A series of carefully designed perceptual tests to evaluate speech quality, emotion identification rates and emotional strength were used for the six emotions which we recorded – *happiness*, *sadness*, *anger*, *surprise*, *fear*, *disgust*. For the HMM-based method, we evaluated spectral and source components separately and identified which components contribute to which emotion.

Our analysis shows that, although the HMM method produces significantly better neutral speech, the two methods produce emotional speech of similar quality, except for emotions having context-dependent prosodic patterns. Whilst synthetic speech produced using the unit selection method has better emotional strength scores than the HMM-based method, the HMM-based method has the ability to manipulate the emotional strength. For emotions that are characterized by both spectral and prosodic components, synthetic speech using unit selection methods was more accurately identified by listeners. For emotions mainly characterized by prosodic components, HMM-based synthetic speech was more accurately identified. This finding differs from previous results regarding listener judgements of speaker similarity for neutral speech. We conclude that unit selection methods require improvements to prosodic modeling and that HMM-based methods require improvements to spectral modeling for emotional speech. Certain emotions cannot be reproduced well by either method.

Key words: Emotional speech synthesis, HMM-based synthesis, unit selection

1. Introduction

The recognition of emotion from human speech, and the generation of synthetic speech that conveys emotion, have a number of applications. For example, in spoken dialog systems, it would be desirable for the system to be able to detect a user's emotional state, alter its subsequent actions, and express an appropriate emotion in its spoken response.

In state-of-the-art TTS methods, such as unit selection (Black and Cambpbell, 1995; Hunt and Black, 1996; Donovan and Woodland, 1999; Syrdal *et al.*, 2000; Clark *et al.*, 2007b) or statistical parametric speech synthesis (Yoshimura *et al.*, 1999, 2000; Zen *et al.*, 2007a, 2009; Yamagishi *et al.*, 2009), reasonably high quality synthetic speech can be produced (Karaiskos *et al.*, 2008), especially for normal neutral reading styles. For example, statistical parametric speech synthesis systems have been found to be as intelligible as human speech (Yamagishi *et al.*, 2008). Concatenative unit selection speech synthesis systems are generally found to produce speech that sounds more similar to the target speaker than statistical parametric speech synthesis systems do (Karaiskos *et al.*, 2008).

Each method has pros and cons for the generation of emotional speech. The main drawback of concatenative methods such as unit selection is that the technique requires a large speech database (e.g., tens or hundreds of hours of speech). To build a system capable of generating emotional speech would require a large database for each of an immense variety of emotions (Bulut *et al.*, 2002; Eide, 2002; Black, 2003; Pitrelli *et al.*, 2006), since this method cannot generalise or interpolate emotions. This would be expensive. To work around this problem, some researchers have attempted to incorporate prosodic or phonologic strategies into unit selection (Hamza *et al.*, 2004; Pitrelli *et al.*, 2006; Hofer *et al.*, 2005), in which rules, found from small or blended emotional speech corpora, are used to modify the target F0 and duration contours (Schröder, 2001). However, this heuristic approach does not always enable the production of emotions for ar-

Email address: barra@die.upm.es (Roberto Barra-Chicote)
[1] Corresponding author

bitrary speakers and the design of an appropriate target cost function is far from easy because the relationship between the components of the target cost and listeners perceptions is unclear (Strom and King, 2008). In addition, since some required subword units may not be present in the small or mixed emotion corpus, this approach may require signal manipulation, which can result in reduced quality synthetic speech. Another practical, but equally severe issue is the accuracy of time-aligned labels of subword units such as diphones or demi-phones. In expressive or emotional speech, more precise labeling than can be obtained from automatic labeling is required (Charonnat *et al.*, 2008; Gallardo-Antolin *et al.*, 2007).

The main drawback of statistical parametric speech synthesis is that the spectra and prosody generated from HMMs tend to be over-smooth and lacking the richness of detail present in natural spectral and prosodic patterns because of the averaging inherent in the statistical approach; these aspects of speech are probably crucial in conveying emotion. In addition, listeners tend to associate negative emotions with robotic voices containing artefacts (Barra *et al.*, 2007). On the other hand, the statistical approach does have notable advantages over unit selection: since all acoustic parameters are modelled within a single framework, it is straightforward to transform or modify the speaking style or emotion by using HMM interpolation (Tachibana *et al.*, 2005), multiple regression of emotion vectors (Nose *et al.*, 2007) and/or HMM adaptation techniques (Tachibana *et al.*, 2006). Since the HMM-based approach requires less data than unit selection and since its flexible voice transformation framework enables the generation of intermediate strengths of emotions or even mixtures of emotions, the HMM-based approach is apparently less costly. Since time-aligned labels are used only for initialization of HMM parameters, the method is far less sensitive to the accuracy of the labels than unit selection.

The annual Blizzard Challenges, run since 2005, provide a clear picture of the performance of various corpus-based speech synthesis techniques (e.g., concatenative, HMM-based or hybrid) for a normal neutral reading style (Black and Tokuda, 2005; Bennett and Black, 2006; Fraser and King, 2007; Karaiskos *et al.*, 2008). However, it is not well understood how well these approaches work for emotional or expressive speech.

One of the major problems in extending the investigation of speech synthesis to emotional speech is that of data collection. As pointed out in (Burkhardt *et al.*, 2005), the so-called "full-blown" emotions very rarely appear in the real world and even if we can obtain recordings of them, there are significant ethical and privacy problems in using such data. Collecting real data is therefore difficult, even more so if we wish to do it in a recording studio. An additional problem is the categorization of emotions: the emotions may be treated as discrete classes or may be treated as continuous variables based on, for example, a Pleasure-Arousal-Dominance model (Schröder, 2004). The socio-cultural background of each individual listener may also cause varying perceptions of emotions, even from the same utterance. Therefore, we chose to use distinct acted emotions for our investigation.

In order to better understand the ability of unit selection and HMM-based methods to produce emotional speech, we built a number of voices from a common corpus, then evaluated them in a series of perceptual tests. The *Spanish Expressive Voices* (SEV) corpus (Barra-Chicote *et al.*, 2008b) was used to build six emotional voices using per method. Synthetic speech generated from each emotional voice was evaluated using from three perspectives: measuring the speech quality; measuring the ability of listeners to correctly identify a given simulated emotion in terms of an emotion identification rate; and measuring the emotional strength of the generated speech.

The paper is organized as follows: in Section 2 the process of building the emotional synthetic voices is described, with details of the corpus and how it was assessed prior to its use in speech synthesis. Section 3 describes the design of the evaluation process, including the evaluation metrics used and the evaluation scenarios. In Sections 4 and 5 the evaluation results are presented and discussed. Section 6 summarises the main findings and identifies outstanding issues for future work.

2. Building Emotional Voices

In this section we describe the SEV emotional speech corpus and the initial subjective evaluation that was carried out to validate the corpus and to assess its quality.

2.1. The SEV Corpus

The *Spanish Expressive Voices* (SEV) corpus (Barra-Chicote *et al.*, 2008b) comprises speech and video recordings of an actor and an actress speaking in a neutral style and simulating six basic emotions: *happiness*, *sadness*, *anger*, *surprise*, *fear* and *disgust*.

Due to a careful design and the relatively large size for a corpus of this type (more than 100 minutes of speech duration per emotion), the recorded database allows comprehensive studies of emotional speech synthesis, prosodic modelling, speech conversion, far-field speech recognition and speech and video-based emotion identification.

The SEV corpus covers speech data in several genres such as isolated word pronunciations, short and long sentences selected from the SES corpus (Montero *et al.*, 1998), narrative texts chosen from a novel "Don Quijote de la Mancha", a political speech given by Spanish philosopher Ortega y Gasset, short and long interviews, question answering situations and short dialogs. The texts of all utterances are emotionally neutral. The structure of the SEV corpus is summarized in Table 1, where the average utterance length of each set is presented in terms of the number of word tokens and number of allophones tokens.

Table 1: Details of the SEV corpus used for building both unit-selection and HMM-based speech synthesis systems. #Utterances represents the number of utterances per emotion. #Words and #Allophones show the average number of word and allophone tokens per utterance, for each subset.

SUBSET	#Utterances	Duration (min)	#Words	#Allophones
Isolated words	570	11	-	-
Short sentences from tales	45	2	5	21
Long sentences from tales	84	8	15	65
Novel	100	11	16	70
A speech	25	4	26	125
Interview (short answers)	52	4	10	44
Interview (long answers)	40	5	20	87
Question-answering	117	4	4	19
Short dialogs	142	5	4	22

The database has been labelled automatically. Phoneme segmentation was performed using HMM forced alignment. Phrase boundaries were determined from the results of the forced alignment. Accentual and syllabic information were predicted from the text. Utterance-medial pauses were inserted into the label sequence according to the prompt text, since the speakers were asked to insert pauses at certain word breaks marked in the prompt.

Finally, the whole database has been validated through objective and perceptual tests, achieving a validation score (emotion identification rate by listeners) as high as 89% (Barra-Chicote *et al.*, 2008b).

Table 2: Percentage of correctly identified utterances for **H**appiness, **A**nger, **Su**rprise, **S**adness, **F**ear, and **D**isgust and the overall average (**AVG**) for the female speaker from the SEV corpus. EIR is the emotion identification rate. $\text{EIR}_{>med}$ is the EIR calculated using only utterances for which the Emotional Strength was rated as *high* or *very high*. $\text{EIR}_{<high}$ is the EIR calculated using only utterances for which the Emotional Strength was lower than *high*.

	H	A	Su	S	F	D	N	AVG
EIR	73	99	76	98	92	97	93	90
$\text{EIR}_{>med}$	76	99	85	98	97	98	98	93
$\text{EIR}_{<high}$	54	96	54	95	75	93	74	77

2.2. Assessment of the SEV Corpus

A further evaluation of the female voice of the corpus was carried out to assess its quality. The evaluation was done by six subjects, three of whom were already familiar with the corpus[2] and three who had never previously heard speech from the corpus. A total set of $3,890$ utterances were presented in random order (approximately 50 minutes per emotion), and at least two listeners evaluated each one. The listeners were required to identify the intended emotion of every given utterance, choosing a label from the six emotions plus an *other* label. Evaluators could listen to each utterance as many times as they needed before providing an answer. Additionally, they were requested to rate the Emotional Strength (ES) of each utterance using a 5-point scale: *very low*, *low*, *medium*, *high* or *very high*.

Table 2 shows the Emotion Identification Rate (EIR) for each emotion and the whole corpus. The EIR was calculated using majority voting between listeners (specifically, only utterances with a full agreement among listeners were considered as correctly identified). $\text{EIR}_{>med}$ is the EIR computed on only those utterances whose ES rank is higher than medium. $\text{EIR}_{<high}$ is the EIR computed on only those utterances whose ES rank is lower than high. Every evaluator converged to a relatively constant EIR for each emotion after about two hundred utterances. The average EIR was 90% and the worst rate for any individual emotion was 73% (for *happiness*). As we might expect, those utterances whose ES rank is higher than medium were easier to identify and thus their EIR was 93%. In contrast, $\text{EIR}_{<high}$ was just 77%. Note that this is distinct emotional speech acted by a professional actress, yet the the emotion identification task is not trivial, even for humans. Specifically, confusions between *happiness* and *surprise* were frequently observed, especially in the short dialog and the question-answering subsets. We also detected that the actress's performance for disgust exhibited a variety of patterns, which was not a problem for trained listeners, but could be difficult for non-trained listeners to cope with. Regarding the emotional strength (ES) evaluation, only 60% of utterances on average were both correctly identified *and* ranked as medium strength or higher.

The results for listeners who were not familiar with the corpus initially contrasted with those of the evaluators, who were familiar with it. However, the unfamiliar listeners appeared to learn and adapt to the evaluation task during the evaluation. Thus, the utterances presented in the early stages of the assessment and the last stages have differing accuracies: listeners gradually improved their identification rates. This adaptation (which we refer to as a

[2]They carried out a partial hand-labelling task of phone durations and pitch marks of a small subset of the SEV corpus. The manually annotated labels were used for other experiments and were not used for this experiment.

"training bias") would not occur if the evaluators or listeners only heard limited numbers of utterances; in that case, all EIR results would be lower, as can be clearly seen in the results described in section 4.2.

2.3. Emotional Voice Building from the SEV corpus

Emotional voices were built using two state-of-the-art synthesis techniques: unit selection and HMM-based synthesis. All voices built are emotion-dependent ones; that is, each emotional voice has been built from scratch using speech data only of the target emotion. The same Spanish text processing modules were used in both synthesis techniques (Barra-Chicote *et al.*, 2008a). The unit selection voices were built using the Multisyn module of the Festival system (Clark *et al.*, 2006, 2007b), and the HMM-based voices were built using a method similar to the Nitech-HTS 2005 system (Zen *et al.*, 2007a) which is publicly available from the HTS toolkit website (Tokuda *et al.*, 2008).

The HMM-based speech synthesis system comprises three components: speech analysis, HMM training, and speech generation. In the speech analysis part, three kinds of parameters for the STRAIGHT (Kawahara *et al.*, 1999) mel-cepstral vocoder with mixed excitation (the mel-cepstrum, $\log F0$ and a set of aperiodicity measures) are extracted as feature vectors for modelling by the HMMs. These are as described in (Zen *et al.*, 2007a), except that the F0 values we used were more robustly estimated using a vote amongst several F0 extraction algorithms (Yamagishi *et al.*, 2009). In the HMM training part, context-dependent multi-stream left-to-right MSD-HSMMs (Zen *et al.*, 2007b) are trained for each emotion using the maximum likelihood criterion. In the speech generation part, acoustic feature parameters are generated from the MSD-HSMMs using the GV parameter generation algorithm (Toda and Tokuda, 2007). Finally an excitation signal is generated using mixed excitation (pulse plus band-filtered noise components) and pitch-synchronous overlap and add (PSOLA) (Moulines and Charpentier, 1990). This signal is used to excite a mel-logarithmic spectrum approximation (MLSA) filter corresponding to the STRAIGHT mel-cepstral coefficients, generating the speech waveform. A Spanish system that we built using the same method on a different corpus exhibited very good performance in a recent Spanish speech synthesis competition (Barra-Chicote *et al.*, 2008a).

3. Design of the Perceptual Evaluation

3.1. Evaluation Metrics

Typically, neutral synthetic speech is evaluated in terms of intelligibility, similarity to the original speaker and the overall quality (or naturalness). For emotional synthetic speech, identification of intended emotions and strength of expressed emotion should be evaluated as well as the overall quality of synthetic speech. We evaluated the following aspects of emotional synthetic speech:

1. Speech Quality (SQ): Listeners were asked to evaluate the overall quality of given emotional synthetic speech using a 5-point scale where 1 was labelled "muy mala" (very bad), 2 was "mala" (bad), 3 was "acceptable" (acceptable), 4 was "buena" (good) and 5 was "muy buena" (very good).
2. Emotion Identification Rate (EIR): The listeners were asked to identify the intended emotion in the given synthetic speech from a limited set of emotional categories: *happiness*, *anger*, *surprise*, *sadness*, *fear*, *neutral*, *disgust*, or *other*.
3. Emotional Strength (ES): The listeners were asked to assess the emotional strength of the given synthetic speech using a continuous slider. The endpoints of the slider were labelled "very weak" and "very strong".

3.2. Experimental Design

Our goal was to get insights into how emotional speech affects the performance of unit selection and HMM-based speech synthesis methods. We also wished to find out how relevant each component of the synthetic voice (segmental information such as spectral envelope, or supra-segmental information as F0, phone duration or power) is to the perception of emotion. We evaluated natural and vocoded speech as well as neutral and emotional synthetic speech generated by the two methods, in order to establish upper bounds in speech quality and emotion identification rate.

In what we call *mixed-emotions experiment*, we conducted a component-level evaluation, in which we separately evaluated the spectral and source components of the HMM-based speech synthesiser. In HMM-based synthesis, the spectral envelope, F0, and duration components are synchronously and simultaneously trained and modeled, but their Gaussian parameters are statistically independent from each other. Thus this framework allows feature-level partial model substitution, which enables us to evaluate the contribution of each component by constructing a mixed model using components taken from a neutral model and an emotional model, as in Fig. 1.

There are eight possible combinations for the model substitution. We chose a subset of these for our experiments, based on the following criteria:

Segmental versus supra-segmental nature
: We separately substituted the emotional spectral envelope with the neutral one and the emotional prosodic components (except power) with neutral ones. This will reveal the relative contribution of the spectral and prosodic components for the perception of each emotion[3].

[3]The spectral properties of speech may be affected by supra-segmental features as well as segmental features. For example, a speaker's laryngeal setting (e.g. breathy phonation) has relatively long-lasting spectral consequences. However, for the simplification of our discussions, we focus on only the purely segmental features of the spectrum in this experiment.

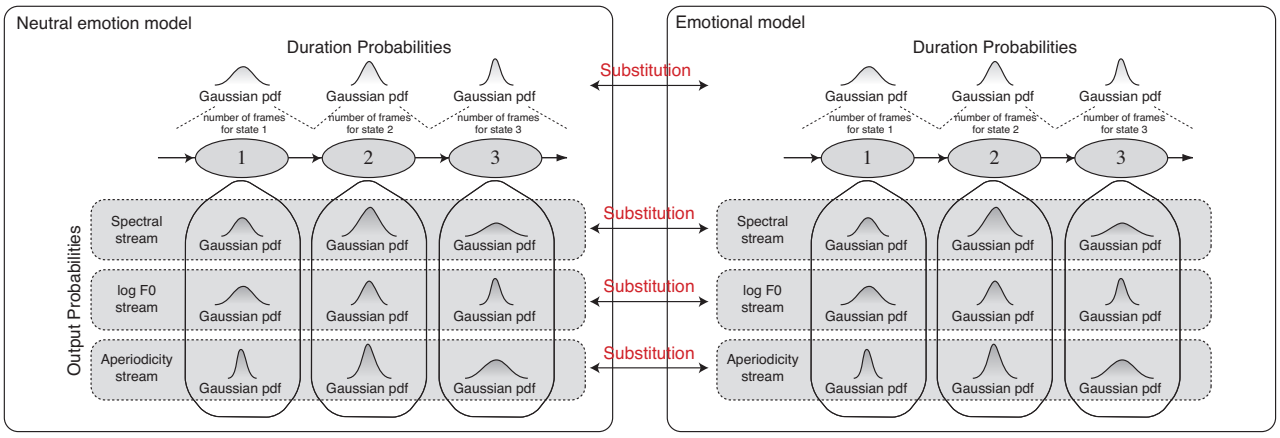

Figure 1: Substitution of each component in an HMM-based speech synthesis system. This enables separate assessment of the importance of spectral and prosody components for emotional speech.

Table 3: Definition and names of speech synthesis systems used for perceptual evaluation. E means emotional, N means neutral, and E_{voc} means vocoded emotional speech.

	SYSTEM		COMPONENTS			PURPOSE
			Spectra	F0 (& Aperiodicity)	Duration	
A	NATURAL SPEECH		E			Evaluation of acted emotions (upper bound of EIR)
B	VOCODED SPEECH		E_{voc}			Evaluation of the impact of vocoder in the EIR
C	UNIT SELECTION	emotional	E			EIR itself and comparison with HMM-BASED
D		neutral	N			
E	HMM	emotional	E			EIR itself and comparison with UNIT SELECTION
F		neutral	N			
G		neutral spectra	N	E	E	spectral versus prosodic nature
H		neutral prosody	E	N	N	
I		neutral F0	E	N	E	Analysis of the relevance of each prosodic feature
J		neutral duration	E	E	N	

Relevance of each prosodic feature

We substituted emotional F0 and duration models with neutral ones. The EIR reduction, relative to the EIR obtained from the "fully emotional" model, will show the contribution of F0 and duration to the perception of each emotion.

3.3. Perceptual Tests and Subjects

The ten systems we built and evaluated are described in Table 3. In our experiments we define a *scheme* as the combination of a synthesis system and a given emotion. A total of 50 *schemes* had to be evaluated (8 systems (A, B, C, E, G, H, I, J in Table 3) combined with six emotions, plus unit selection neutral (D in Table 3) and HMM-based neutral systems (F in Table 3)).

A fully factorial experimental design would require each listener to hear too many stimuli, which has a number of drawbacks with regard to the evaluators' limited attention span and the "training bias" problem we described in Section 2.2. Instead, the experimental design was based on a balanced latin-square matrix (Gomes *et al.*, 2004), similar to the experimental design used in the Blizzard Challenge (Fraser and King, 2007).

Each *scheme* generated a set of speech for the fifty sentences that were not included in the voice training sentences. They were medium length sentences, between 6 to 17 words and with an average length of 10 words. The content of the test sentences was emotionally neutral to allow listeners to focus on acoustic cues only. The latin-square design allow evaluation of all schemes and all synthesized sentences and controls for ordering effects by ensuring that each group of listeners hears the stimulii in a different order.

Fifty listeners, having a similar socio-cultural profile, participated in the evaluation, which was carried out individually in a single session per listener. All listeners were from the Madrid area and were between twenty and forty years old, and none of them had a speech-related research background nor had they previously heard any of the SEV speech recordings. The evaluation was conducted in a quiet environment using headphones. The authors decided to avoid long sessions, thus limiting to 50 the number of stimuli to be presented to each listener, so that the av-

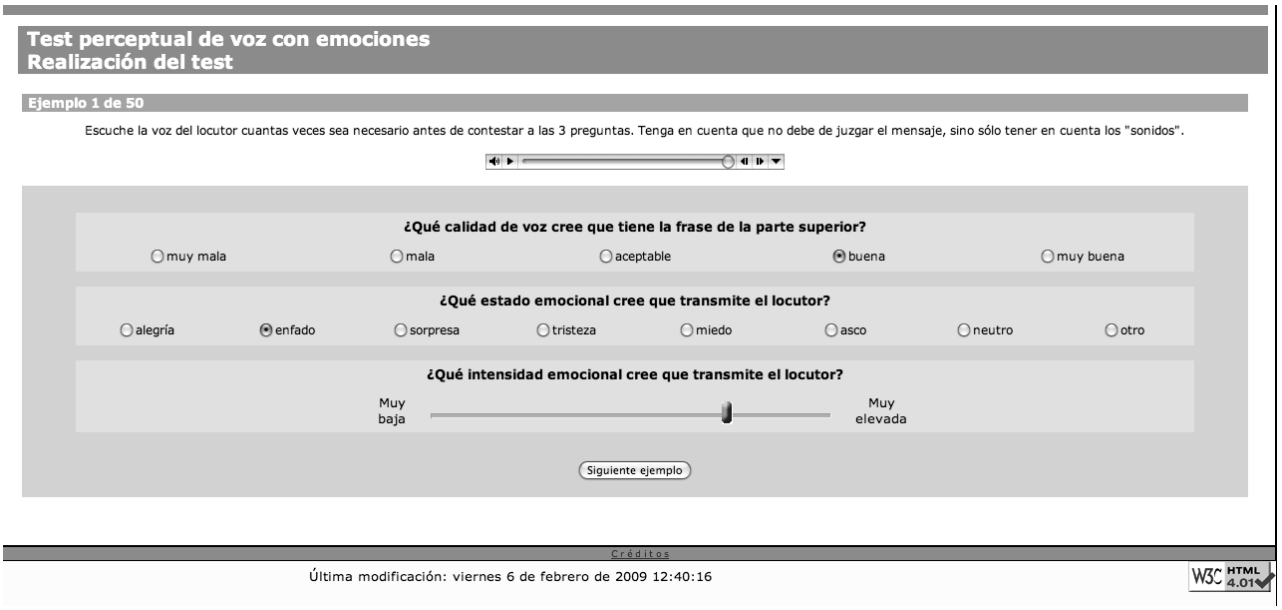

Figure 2: Web interface used by the listener for a perceptual evaluation.

erage length of each session was 31 minutes.

The evaluation was conducted via a web browser interface shown in Figure 2. Speech quality, intended emotion and emotional strength were all evaluated in the same trial (and in this same "visual" order in the web page). Note that listeners were explicitly asked to make each judgement independently from the others. Before making a decision, each utterance could be played as many times as the listener wished, but they could never go back to re-evaluate previous utterances.

The evaluation using 50 listeners provided total of 300 evaluation responses (i.e. 50 per emotion) for each system, except for systems D and F in Table 3. Systems D and F are for the neutral emotion and have 50 evaluation responses for each. One evaluation response includes the listener's rating for speech quality, the identified emotion and the emotional strength, for a single stimulus.

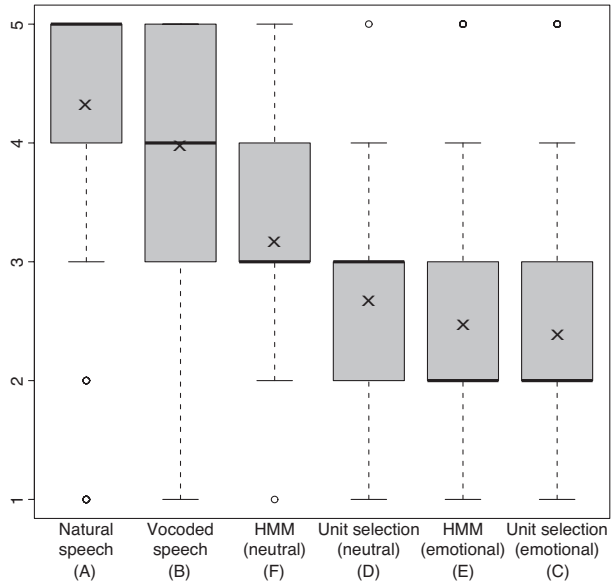

Figure 3: Boxplot showing speech quality (SQ) scores for natural emotional speech, vocoded natural emotional speech, unit selection synthetic speech and HMM-based synthetic speech. For the emotional voices, the result shown is the average over the six emotions.

4. Comparison of Speech Synthesis Methods

4.1. Speech Quality (SQ)

Figure 3 presents a boxplot showing SQ results for the speech synthesis methods, where the median is represented by a solid bar across a box showing the quartiles with whiskers extending to 1.5 times the inter-quartile range and outliers beyond this being represented as circles. The mean is represented by a cross. As explained in (Clark *et al.*, 2007a), SQ scores are ordinal data and therefore we used a series of pairwise Wilcoxon signed rank tests with Bonferroni step-down correction to determine whether there are significant differences between the SQ scores of systems. Table 4 shows the significant differences between systems at $p = 0.05$.

The SQ scores for natural speech and vocoded speech are high, as expected. For neutral speech, the HMM-based method (system F) has significantly better quality than that of unit selection (system D) whereas for emotional speech there is no significant difference between them. We can also see that emotional synthetic voices using both methods have worse quality than the neutral synthetic voices. Since all the neutral and emotional voices were trained on the same amount of speech data uttered by the same speaker, the results demonstrate that emotional

Table 4: Results of pairwise Wilcoxon signed rank tests between natural speech (A), vocoded speech (B), unit selection (emotional) (C), unit selection (neutral) (D), HMM-based (emotional) (E) and HMM-based (neutral) (F). ■ denotes a significant difference in Speech Quality (SQ) between a pair of systems (significance level is $p = 0.05$).

	A	B	C	D	E
B	■				
C	■	■			
D	■	■			
E	■	■			
F	■	■	■	■	■

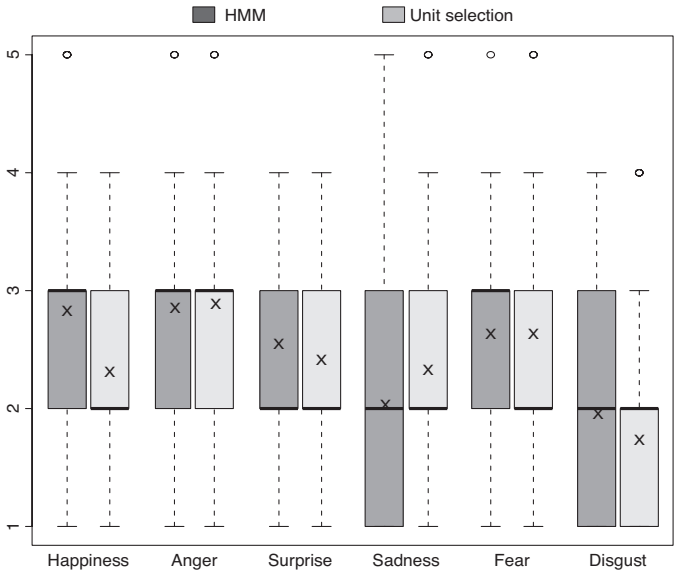

Figure 4: Boxplot showing speech quality (SQ) scores for unit selection synthetic speech and HMM-based synthetic speech for each emotion.

Table 5: Results of Wilcoxon signed rank tests with Bonferroni step-down correction for SQ scores between emotions. ■ denotes a significant difference in SQ between emotions at a significance level of $p = 0.05$. The first row and column give the emotions: **H**appiness, **A**nger, **Su**rprise, **S**adness, **F**ear and **D**isgust.

	H	A	Su	S	F
A					
Su		■			
S	■	■	■		
F				■	
D	■	■	■		■

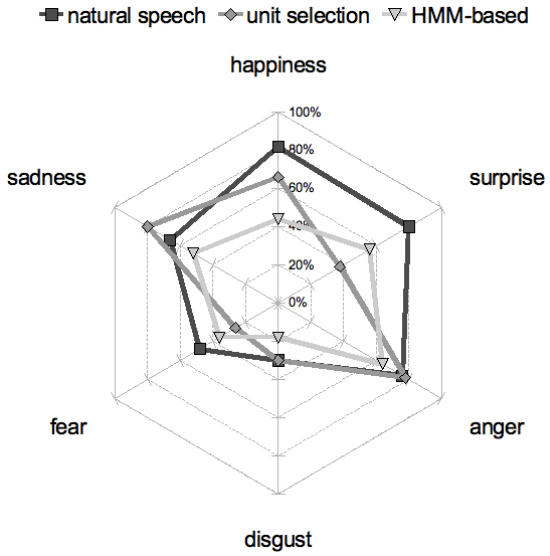

Figure 5: EIRs for natural emotional speech, unit selection synthetic emotional speech and HMM-based synthetic emotional speech.

speech utterances are more difficult to model.

Figure 4 presents the boxplot showing the SQ scores for each emotion using HMM-based and unit selection methods. Table 5 shows the significant differences between emotions at a significance level of $p = 0.05$. The emotional voice with the highest score for both techniques is *anger*. The average score for both the methods for *anger* is significantly better than the scores for *surprise*, *sadness*, and *disgust*. On the other hand, *disgust* has the lowest score for both techniques. Its average score is significantly worse than all the other emotions except *sadness*. These differences between emotions, regardless of the synthesis method used, deserves further investigation and will be the subject of future work.

Pairwise Wilcoxon signed rank tests were also conducted between HMM-based and unit selection systems for each emotion. Only for *happiness*, is there a statistically significant difference (at $p < 0.05$) between the SQ score for the HMM-based and unit selection methods. We presume that unit selection method has difficulty coping with the wider variety of happy-speech prosodic patterns within similar linguistic contexts (Scherer, 2003), because the evaluation of the cost functions is based on local features and there is no prosody modification in the synthesis process. On the contrary, because HMM-based synthesis explicitly models the prosody using wide-context supra-segmental features, it is able to generate smoother prosodic patterns. However, it appears that this context-dependency issue is less important for *fear* because the prosodic patterns are more uniform over all utterances (Schröder, 2001).

4.2. Emotion Identification Rates (EIR)

The EIR results are shown in Figure 5. The average EIR for natural speech is 64%. *Surprise*, *happiness*, and *anger* are correctly identified over 70% of the time. *Sadness* is identified over 60% of the time and *fear* is identified 50% of the time. *Disgust* was least well identified, with an EIR of just 30%.

The unit selection method has better absolute EIRs for *happiness*, *anger*, *sadness*, and *disgust* and the HMM-based method has better absolute EIRs for *surprise* and

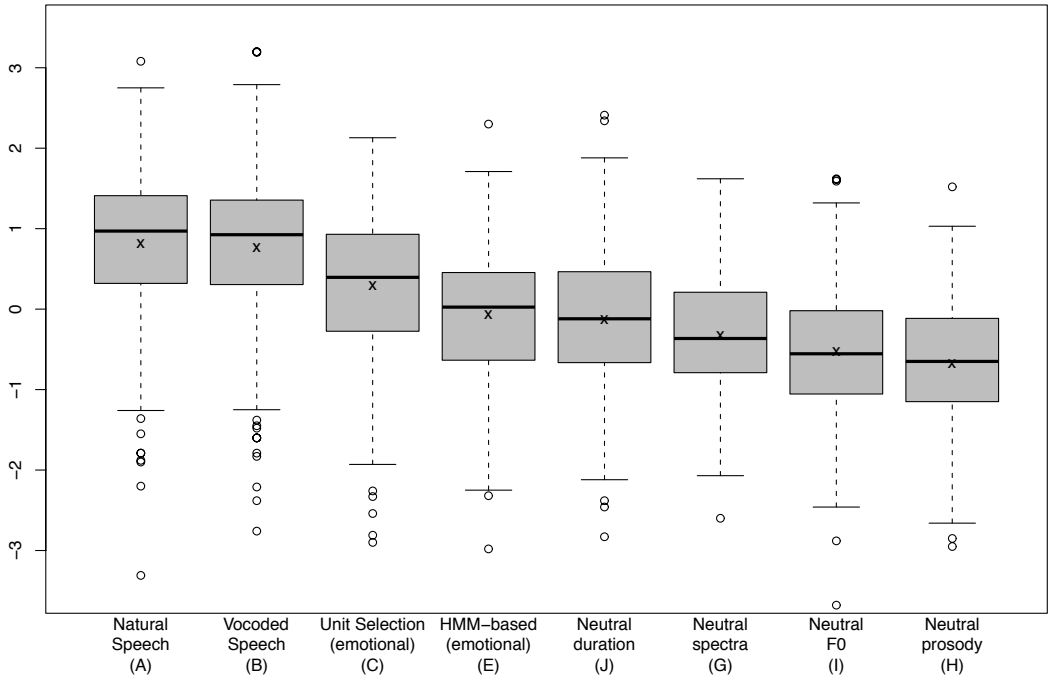

Figure 6: Boxplot showing the normalized emotional strength (ES) scores obtained for each system.

Table 6: EIR relative to natural speech EIR, expressed as a percentage. The first row shows the emotions: **H**appiness, **A**nger, **Su**rprise, **S**adness, **F**ear and **D**isgust.

	H	A	Su	S	F	D	AVG
Unit selection	80	103	48	121	54	100	84
HMM-based	54	84	70	79	75	60	70

Table 7: Results of pairwise t-tests over normalized ES scores and between natural speech (A), vocoded speech (B), unit selection (emotional) (C), HMM-based (emotional) (E), neutral spectra (G), neutral prosody (H), neutral F0 (I) and neutral duration (J). ■ denotes a significant difference in ES between systems (at $p = 0.05$).

	A	B	C	E	G	H	I
B							
C	■	■					
E	■	■	■				
G	■	■	■	■			
H	■	■	■	■	■		
I	■	■	■	■	■		
J	■	■	■			■	■

fear. This is interesting because, in neutral speech, unit selection systems generally achieve better speaker similarity scores than HMM-based systems (Karaiskos *et al.*, 2008). We analyze this result in Section 4.3. For the *disgust* emotion, both methods have their lowest EIR (18% for HMM and 30% for unit selection) and low SQ scores. The low EIR also obtained using natural speech for this emotion indicates that this emotion, at least as portrayed by the actress, is hard to identify.

Table 6 shows the EIR of synthetic speech, relative to the EIR of natural speech. This illustrates the discrepancies between identification of synthetic emotional speech and natural emotional speech. First, we can see that unit selection voices are generally more accurately identified than HMM-based voices. For individual emotions we can see that both methods have higher relative EIRs for *anger* and *sadness*, indicating that these emotions can be detected in synthetic speech about as well as in natural speech. On the other hand, we see that the unit selection method has a very low relative EIR for *surprise* and the HMM-based method has relatively a low relative EIR for *happiness*. One of the main features of *surprise* as acted by the actress is a raising of F0 in the last stress group in an utterance. Neither method could reproduce this phenomenon well, particularly unit selection.

The unit selection voice for *sadness* obtains a higher EIR than natural speech. In our previous experiments on phonetic segmentation of this emotional speech corpus, we observed that many segmentation errors were due to parts of relatively long pauses sometimes being labelled as a part of an adjacent phoneme. This suggests that a cause for this unusually high EIR for unit selection speech is the insertion of "false pauses" that are interpreted by listeners as *sadness*.

4.3. Emotional Strength (ES)

The Emotional Strength (ES) score, elicited from the listeners using a slider, was treated as a continuous variable

without categorical information. Since every listener may use his or her own scale, we normalized the scores on a per listener basis.

The boxplot presented in Figure 6 shows the normalized ES scores. We performed an ANOVA test to evaluate which factors had a statistically significant influence on the normalized ES scores. The ANOVA test results showed that the effect of system is strongly significant ($p < 2.2e-16$). Given this result, Pairwise t-tests with Bonferroni step-down correction were conducted to determine whether there are significant differences between the normalized ES scores of each system. Table 7 shows the significant differences between systems for $p = 0.05$.

As expected, natural and vocoded speech (systems A and B, respectively) obtained the highest ES scores. The unit selection method (system C) has a better overall ES score than the HMM-based method (system E). We assume that the statistical averaging process inherent in the HMM modeling process causes emotional strength to become weaker, in the same way that similarity to the original speaker is also somewhat reduced. However it is, in theory, possible for the HMM-based method to emphasize an emotion and so it could potentially improve the emotional strength, by using extrapolation (Tachibana *et al.*, 2005). Although we did not do this in the experiments here, interpolated and extrapolated examples for *anger* are available from `http://lorien.die.upm.es/~barra/voices/emotions-interpolation` and English samples are available from `http://homepages.inf.ed.ac.uk/jyamagis/Demo-html/Style.html`.

5. Feature-level analysis for the HMM-based method

The mixed emotion systems were evaluated as part of the same listening test described in the previous section, but we present the analysis separately in this section. The SQ scores, EIRs, and ES scores for the mixed-emotion systems are shown in Figure 7, Table 9 and Figure 6, respectively. Table 8 shows the significant differences found between the systems whose results are given in Figure 7. The EIRs are given as a percentage relative to the EIR of the corresponding fully emotional HMM-based system.

From Figure 7, we can see that mixing emotional models and neutral models generally results in a reduction in the SQ score (see "neutral duration", "neutral spectra", "neutral F0" and "neutral prosody"). This effect is most apparent for the neutral prosody system, which suggest that the F0 and duration parameters significantly contribute to the naturalness of emotional synthetic speech. This was also confirmed with pairwise Wilcoxon signed rank tests, in which statistically significant differences were only found between the emotional HMM-based system and the neutral prosody system.

From the results in Table 9, we can distinguish a clear pattern of either a high relative EIR or a very low relative EIR, with only a few intermediate values. If we say that an

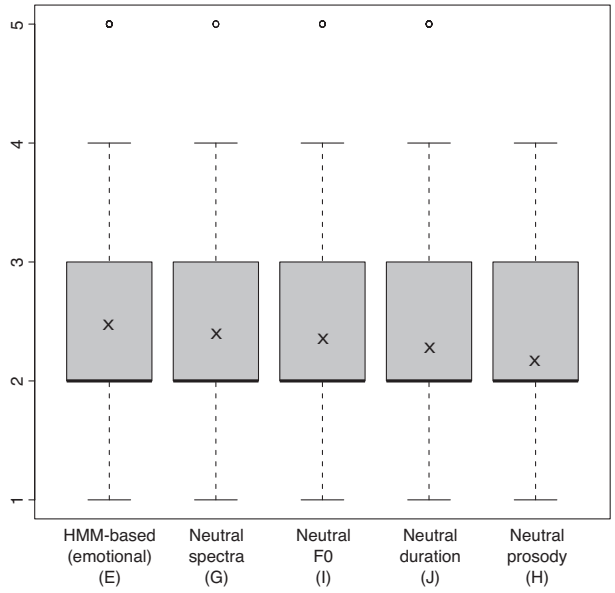

Figure 7: Boxplot showing speech quality (SQ) scores for HMM-based synthetic speech, neutral spectra, neutral prosody, neutral F0 and neutral duration for each emotion.

Table 8: Results of pairwise Wilcoxon signed rank tests between HMM-based (emotional) (E), neutral spectra (G), neutral prosody (H), neutral F0 (I) and neutral duration (J). ■ shows a significant difference in SQ between systems (at $p = 0.05$).

	E	G	H	I
G				
H	■			
I				
J				

EIR reduction of 75% or more (i.e., a relative EIR value of 25% or less in the table) indicates that those model components are essential for reproduction of that, we may conclude that:

1. *Happiness* and *disgust* are characterized by both spectral and prosodic components.
2. *Anger* is mainly characterized by the spectral components. Note that the spectral components include power coefficients as well as coefficients for spectral envelope.
3. *Surprise* and *fear* are mainly characterized by the prosodic components.
4. *Sadness* cannot be characterized by either spectral or prosodic components well. The frequency of pause may affect this emotion.

Note that the HMM-based method had better absolute EIRs for *surprise* and *fear* in Section 4.2. Since these emotions are mainly characterized by the prosodic components, we conclude that the HMM-based method has good prosodic modeling. In turn, for emotions that are characterized by not only the prosodic components but also the

Table 9: EIR for mixed-emotion systems, relative to the EIR for a fully emotional HMM-based system, expressed as a percentage. The first row shows the emotions: **H**appiness, **A**nger, **Su**rprise, **S**adness, **F**ear, and **D**isgust. Underlined values highlight EIRs below 25%.

	H	A	Su	S	F	D
Emotional	100	100	100	100	100	100
Neutral spectra	18	13	75	81	72	22
Neutral prosody	27	59	4	58	6	22
Neutral F0	9	91	0	62	11	89
Neutral duration	64	59	82	150	106	33

spectral components, synthetic speech using unit selection method is more accurately identified by listeners, because the spectral modeling of HMM-based method is not as good as the "spectral modelling" (i.e., unit playback) of the unit selection method. This is most likely due to the over-smooth spectra generated by the HMMs, which is an inherent and unsolved problem of the statistical approach.

From Figure 6 we can see that, compared to fully emotional HMM voices, the ES score decreases in the following order:

1. neutral duration
2. neutral spectra
3. neutral F0
4. neutral prosody (neutral F0 + neutral duration)

This means that F0 is the most relevant parameter for emotional strength and that the prosodic components (F0 and duration, system H) are significantly more relevant to emotional strength than the spectral envelope (system G) or duration (system J). In other words, accurate models for some supra-segmental features such as F0 and duration are more important than spectral envelope features in emotional speech. These results are also consistent with our previous analysis (Barra *et al.*, 2007) and results from automatic emotion identification tasks (Barra *et al.*, 2006).

6. Conclusions and Future Work

Two state-of-the-art speech synthesis techniques (unit selection and HMM-based synthesis) were applied to emotional speech. A series of perceptual tests were used to evaluate voices built for six emotions available in the SEV corpus: *happiness*, *sadness*, *anger*, *surprise*, *fear*, and *disgust*. For the HMM-based method, we also evaluated spectral and source components separately and identified which components contribute to the reproduction of each emotion.

Although the HMM method produced significantly better neutral speech, synthetic emotional speech generated from HMMs and from unit selection has comparable speech quality, except for emotions having context-dependent prosodic patterns. Synthetic speech produced using unit selection has better emotional strength scores than when using the HMM method. The HMM-based method does have an ability to manipulate the emotional strength, although this was not explored in these experiments. For emotions that are characterized by both spectral and prosodic components, synthetic speech using unit selection is more accurately identified. When emotions are mainly characterized by the prosodic components, HMM-based synthetic speech is more accurately identified. These results suggests that unit selection methods require improvements in prosodic modeling whereas HMM-based methods require improvements in spectral modeling, for the synthesis of emotional speech.

Neither technique could reproduce the *disgust* emotion well, although listeners' identification rates even on natural speech were low for this emotion. The exact causes of this are a subject for future investigation.

The unit selection and HMM-based synthesis systems built are available at the SEV online demonstration page: `http://lorien.die.upm.es/~barra/voices/index.php?status=voices`

Acknowledgements

RB was visiting CSTR at the time of this work. RB was supported by the Spanish Ministry of Education and by project ROBONAUTA (DPI2007-66846-c02-02). JY is supported by EPSRC and the EC FP7 EMIME project. SK holds an EPSRC Advanced Research Fellowship. JMM and JMG are supported by projects SD-TEAM-UPM (TIN2008-06856-C05-03) and SD-TEAM-UAH (TIN2008-06856-C05-05), respectively. This work has made use of the resources provided by the Edinburgh Compute and Data Facility which is partially supported by the eDIKT initiative (http://www.edikt.org.uk). We also thank the two anonymous reviewers for their constructive feedback and helpful suggestions. The associate editor coordinating the review of this manuscript for publication was Dr. Marc Swerts.

References

Barra, R., Montero, J., Macias-Guarasa, J., D'Haro, L., San-Segundo, R., and Cordoba, R. (2006). Prosodic and segmental rubrics in emotion identification. In *ICASSP 2006*, pages 1085–1088.

Barra, R., Montero, J., Macias-Guarasaa, J., Gutierrez-Arriola, J., Ferreiros, J., and Pardo, J. (2007). On the limitations of voice conversion techniques in emotion identication tasks. In *Proc. Interspeech 2007*, pages 2233–2236.

Barra-Chicote, R., Yamagishi, J., Montero, J., King, S., Lutfi, S., and Macias-Guarasa, J. (2008a). Generacion de una voz sintetica en Castellano basada en HSMM para la Evaluacion Albayzin 2008: conversion texto a voz. In *V Jornadas en Tecnologia del Habla*, pages 115–118. (in Spanish).

Barra-Chicote, R., Montero, J., Macias-Guarasa, J., Lufti, S., Lucas, J. M., Fernandez, F., D'haro, L., San-Segundo, R., Ferreiros, J., Cordoba, R., and Pardo, J. (2008b). Spanish Expressive Voices: Corpus for emotion research in Spanish. In *Proc. of 6th international conference on Language Resources and Evaluation*.

Bennett, C. and Black, A. W. (2006). The Blizzard Challenge 2006. In *Proc. Blizzard Challenge 2006*.

Black, A. W. (2003). Unit selection and emotional speech. In *Proc. EUROSPEECH 2003*, pages 1649–1652.

Black, A. W. and Cambpbell, N. (1995). Optimising selection of units from speech database for concatenative synthesis. In *Proc. EUROSPEECH-95*, pages 581–584.

Black, A. W. and Tokuda, K. (2005). The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets. In *Proc. EUROSPEECH 2005*, pages 77–80.

Bulut, M., Narayan, S., and Syrdal, A. (2002). Expressive speech synthesis using a concatenative synthesizer. In *Proc. ICSLP 2002*, pages 1265–1268.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005). A database of German emotional speech. In *Proc. Interspeech 2005*, pages 1517–1520.

Charonnat, L., Vidal, G., and Boeffard, O. (2008). Automatic phone segmentation of expressive speech. In *Proc. Language Resources and Evaluation Conference*, pages 2376–2379.

Clark, R., Richmond, K., Strom, V., and King, S. (2006). Multisyn voice for the Blizzard Challenge 2006. In *Proc. Blizzard Challenge Workshop 2006*.

Clark, R., Podsiadlo, M., Fraser, M., Mayo, C., and King, S. (2007a). Statistical analysis of the Blizzard Challenge 2007 listening test results. In *Proc. BLZ3-2007 (in Proc. SSW6)*.

Clark, R. A., Richmond, K., and King, S. (2007b). Multisyn: Open-domain unit selection for the Festival speech synthesis system. *Speech Communication*, **49**(4), 317–330.

Donovan, R. and Woodland, P. (1999). A hidden Markov-model-based trainable speech synthesizer. *Computer Speech and Language*, **13**(3), 223–241.

Eide, E. (2002). Preservation, identification, and use of emotion in a text-to-speech system. In *Proc. of IEEE workshop on Speech Synthesis*, pages 127–130.

Fraser, M. and King, S. (2007). The Blizzard Challenge 2007. In *Proc. BLZ3-2007 (in Proc. SSW6)*.

Gallardo-Antolin, A., Barra, R., Schröder, M., Krstulovic, S., and Montero, J. (2007). Automatic phonetic segmentation of Spanish emotional speech. In *Proc. InterSpeech 2007*.

Gomes, C., Sellmann, M., Es, C. V., and Es, H. V. (2004). The challenge of generating spatially balanced scientific experiment designs. In *In CP-AI-OR'04*, pages 387–394.

Hamza, W., Bakis, R., Eide, E., Picheny, M., and Pitrelli, J. (2004). The IBM expressive speech synthesis system. In *Proc. ICSLP 2004*.

Hofer, G., Richmond, K., and Clark, R. (2005). Informed blending of databases for emotional speech synthesis. In *Proc. Interspeech 2005*, pages 501–504.

Hunt, A. and Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. ICASSP-96*, pages 373–376.

Karaiskos, V., King, S., Clark, R. A. J., and Mayo, C. (2008). The Blizzard Challenge 2008. In *Proc. Blizzard Challenge Workshop 2008*, Brisbane, Australia.

Kawahara, H., Masuda-Katsuse, I., and Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. *Speech Communication*, **27**, 187–207.

Montero, J. M., Gutierrez-Arriola, J. M., Palazuelos, S., Enriquez, E., Aguilera, S., and Pardo, J. M. (1998). Emotional speech synthesis: From speech database to TTS. In *Proc. ICSLP-98*, pages 923–926.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, **9**(5-6), 453–468.

Nose, T., Yamagishi, J., and Kobayashi, T. (2007). A style control technique for HMM-based expressive speech synthesis,. *IEICE Trans. Inf. & Syst.*, **E90-D**(9), 1406–1413.

Pitrelli, J., Bakis, R., Eide, E., Fernandez, R., Hamza, W., and Picheny, M. (2006). The IBM expressive text-to-speech synthesis system for American English. *IEEE Trans. on Speech Audio Process.*, **14**(4), 1099–1108.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, **40**(1-2), 227 – 256.

Schröder, M. (2001). Emotional speech synthesis: a review. In *Proc. EUROSPEECH 2001*, pages 561–564.

Schröder, M. (2004). *Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis*. Ph.D. thesis, Saarland University, Saarland.

Strom, V. and King, S. (2008). Investigating Festival's target cost function using perceptual experiments. In *Proc. Interspeech 2008*, pages 1873–1876.

Syrdal, A., Wightman, C., Conkie, A., Stylianou, Y., Beutnagel, M., Schroeter, J., Storm, V., Lee, K., and Makashay, M. (2000). Corpus-based techniques in the AT&T NEXTGEN synthesis system. In *Proc. ICSLP 2000*, pages 411–416.

Tachibana, M., Yamagishi, J., Masuko, T., and Kobayashi, T. (2005). Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing,. *IEICE Trans. Inf. & Syst.*, **E88-D**(11), 2484–2491.

Tachibana, M., Yamagishi, J., Masuko, T., and Kobayashi, T. (2006). A style adaptation technique for speech synthesis using HSMM and suprasegmental features. *IEICE Trans. Inf. & Syst.*, **E89-D**(3), 1092–1099.

Toda, T. and Tokuda, K. (2007). A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Trans. Inf. & Syst.*, **E90-D**(5), 816–824.

Tokuda, K., Zen, H., Yamagishi, J., Masuko, T., Sako, S., Black, A., and Nose, T. (2008). *The HMM-based speech synthesis system (HTS) Version 2.1*. http://hts.sp.nitech.ac.jp/.

Yamagishi, J., Zen, H., Wu, Y.-J., Toda, T., and Tokuda, K. (2008). The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge. In *Proc. Blizzard Challenge 2008*.

Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., King, S., and Renals, S. (2009). A robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Trans. Speech, Audio & Language Process.*, **17**(6), 1208–1230.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. EUROSPEECH-99*, pages 2374–2350.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. *IEICE Trans.*, **J83-D-II**(11), 2099–2107. (in Japanese).

Zen, H., Toda, T., Nakamura, M., and Tokuda, K. (2007a). Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. *IEICE Trans. Inf. & Syst.*, **E90-D**(1), 325–333.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2007b). A hidden semi-Markov model-based speech synthesis system. *IEICE Trans. Inf. & Syst.*, **E90-D**(5), 825–834.

Zen, H., Tokuda, K., and Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, **51**(11), 1039–1064.